НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2024 Том 24 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2024 Vol. 24 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-5-788-796

Enhanced anomaly detection in network security: a comprehensive ensemble approach

Rashmikiran Pandey[1]✉, Mrinal Pandey[2], Alexey N. Nazarov[3]

[1,2] Moscow Institute of Physics and Technology (National Research University), Moscow region, Dolgoprudny, 141701, Russian Federation
[3] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333, Russian Federation

[1] rashmikiran@phystech.edu✉, https://orcid.org/0000-0003-0042-6565
[2] mrinalpandei@phystech.edu, https://orcid.org/0009-0009-5151-6908
[3] a.nazarov06@bk.ru, https://orcid.org/0000-0002-0497-0296

Abstract
Detection and handling of anomalous behavior in the network systems are peremptory efforts to ensure security for vulnerable infrastructures amidst the dynamic context of cybersecurity. In this paper, we propose an ensemble machine learning model architecture that leverages the strengths of XGBoost, Gradient Boosting, Random Forest, and Support Vector Machine models to identify anomalies in the dataset. This method utilizes an ensemble of these models with weighted voting based on accuracy to enhance anomaly detection for robust and adaptive real-world network security. The proposed ensemble learning model is evaluated on standard metrics and demonstrates exceptional efficacy, achieving an impressive accuracy of 99.68 % on NSL KDD dataset. This remarkable performance extends the model prowess in discerning anomalies within network traffic showcasing its potential as a robust tool for enhancing cybersecurity measures against evolving threats.

Keywords
anomaly detection, bagging and boosting, ensemble approach, network security, neural network

For citation: Pandey R., Pandey M., Nazarov A.N. Enhanced anomaly detection in network security: a comprehensive ensemble approach. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 788–796. doi: 10.17586/2226-1494-2024-24-5-788-796

УДК 004.89

Расширенное обнаружение аномалий в сетевой безопасности: комплексный ансамблевый подход

Рашмикиран Пандей[1]✉, Мринал Пандей[2], Алексей Николаевич Назаров[3]

[1,2] Московский физико-технический институт (национальный исследовательский университет), Московская область, Долгопрудный, 141701, Российская Федерация
[3] Федеральный исследовательский центр «Информатика и управление» Российской академии наук», Москва, 119333, Российская Федерация

[1] rashmikiran@phystech.edu✉, https://orcid.org/0000-0003-0042-6565
[2] mrinalpandei@phystech.edu, https://orcid.org/0009-0009-5151-6908
[3] a.nazarov06@bk.ru, https://orcid.org/0000-0002-0497-0296

Аннотация
Обнаружение и устранение аномального поведения сетевых систем являются важнейшими мерами по обеспечению безопасности уязвимых инфраструктур в динамичном контексте кибербезопасности. Предложена архитектура модели машинного обучения ensemble, которая использует преимущества моделей XGBoost, Gradient Boosting, случайного леса и метода опорных векторов для выявления аномалий в наборе данных. Представленный подход использует совокупность перечисленных моделей с взвешенным голосованием и основан на точности, для улучшения обнаружения аномалий и обеспечения надежной и адаптивной сетевой безопасности в реальном времени. Модель коллективного обучения оценивается по стандартным показателям

и демонстрирует исключительную эффективность, достигая высокой точности 99,68 % в наборе данных NSL KDD. Высокая производительность подхода расширяет возможности модели в выявлении аномалий в сетевом трафике, демонстрирует ее потенциал в качестве надежного инструмента для усиления мер кибербезопасности против развивающихся угроз.

Ключевые слова
обнаружение аномалий, пакетирование и бустинг, групповой подход, сетевая безопасность, нейронная сеть

Ссылка для цитирования: Пандей Р., Пандей М., Назаров А.Н. Расширенное обнаружение аномалий в сетевой безопасности: комплексный ансамблевый подход // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 5. С. 788–796 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-5-788-796

Introduction

In the ever-evolving cyber landscape, cybersecurity systems face a formidable challenge in safeguarding networked environments from anomalous activities. These threats, ranging from sophisticated intrusions to stealthy attacks, pose unprecedented threats to the confidentiality, integrity, and availability of sensitive information [1]. While conventional signature-based intrusion detection systems are effective in identifying known malicious threats, they often fall short in detecting novel and previously unseen anomalies. This highlights the urgent need for advanced anomaly detection techniques that can adapt to the ever-changing nature of cyber threats [2].

Existing work have intensely analyzed the applicability of single-classifier machine learning models for anomaly detection. However, failure of the standalone machine learning models to effectively address the complexities of network security has fuelled further investigation in this domain. The network security domain encapsulates a unique sort of challenges like imbalanced datasets evolving methodologies of attack and the requisite for real-time detection [3]. The development of novel methods capable of identifying anomalous patterns in huge network datasets is a requirement that arises from the ever-growing sophistication and diversity of cyber threats. In response, we introduce a bridging-the-gap ensemble framework especially for the network security applications. With the ensemble methods having an advanced framework, our proposed research will venture in this promising avenue with a sophisticated framework combining the prediction abilities of few models to reach higher accuracy, improved robustness, and generalizing ability [4]. The lower-level inbuilt models in our research come up as promising experts along with the ensemble approach. However, prior to the use of ensemble methods in predictive modeling, such research frequently falls short of an effective way to choose those individual models which are built into the ensemble [5–7]. Recognizing this crucial gap, the following proposes to breach this chasm by the introduction of a novel ensemble model that capitalizes on the strengths of well-established algorithms — XGBoost, Random Forest, Gradient Boosting, and Support Vector Machines. So, the basic idea of our approach is to develop a synergy ensemble model that simultaneously combines the strengths of both bagging and boosting techniques to enhance the robustness as well as predictive performances of the anomaly detection system [8]. Our proposed model, utilizing the NSL KDD dataset for in-depth analysis and evaluation, endeavors to achieve three of objectives: firstly, we aspire to cultivate an ensemble model that seamlessly integrates the predictive prowess of diverse machine learning techniques for anomaly detection. Secondly, we ardently pursue the rigorous evaluation and comparative analysis of the proposed ensemble model against individual models, employing established metrics such as accuracy, precision, recall, and F1-score. Thirdly, we eagerly seek to impart profound insights into the effectiveness of ensemble methods in bolstering the robustness and generalization of anomaly detection systems within the intricate domain of network security. The experimental showcase the outstanding accuracy of 99.68 % in proposed ensemble model that demonstrating the efficacy of the model. This remarkable achievement presents a new frontier in the realm of ensemble methods, paving the way for accurate identification and classification of anomalies within network traffic.

Literature Review

The field of network anomaly detection has witnessed significant evolution in response to the growing complexity of cyber threats and the escalating need for robust cybersecurity solutions. This study in [9] proposes a new method for network anomaly detection using a 5-layer Auto Encoder (AE) consisting of 1 input, 1 output followed by 2 dense and a bottleneck layer. The key feature is a data pre-processing step that tackles the issue of data imbalance. The authors' approach combats this by transforming and removing outliers that significantly skew the data. Another innovation lies in the model core: a new mean absolute error based reconstruction error function. This function plays a crucial role in classifying network traffic as normal or anomalous. The paper emphasizes that this function, combined with an optimal 5-layer AE architecture, allows for superior feature learning. By compressing the data into a lower-dimensional space, the model focuses on the most critical characteristics for anomaly detection. The authors tested their model on the NSL KDD dataset and achieved 90.61 % accuracy and 92.26 % F1-score in anomaly detection. Similar to this approach, another study [10] achieved an accuracy of 85 % using an AE model on the same NSL KDD dataset. In contrast, a different approach leveraging deep learning is presented by Bhavna et al. [11]. This study investigates a Convolutional Neural Network (CNN) model for Network Intrusion Detection Systems (NIDS). They also utilize the NSL KDD benchmark dataset, but their model focuses on four specific attack classes. To improve efficiency, they employ a filter-based feature reduction method to remove redundant features within the dataset. The core of their system is a 2D-CNN model, which achieves an impressive accuracy of 99.4 % with

reduced loss. The paper likely compares the performance of both models (AE and CNN) in terms of accuracy and other evaluation metrics providing valuable insights into the effectiveness of different deep learning architectures for NIDS tasks. Yet another deep learning approach is presented in [6], where the authors propose a model that combines the strengths of Bidirectional Long Short-Term Memory (BLSTM) and attention mechanism for anomaly detection. BLSTM is adept at learning the characteristics of network traffic data by analyzing sequential data like network packets. The attention mechanism refines this process by focusing on the most critical features within the BLSTM output, effectively identifying key aspects for traffic classification. However, this model achieves a lower accuracy of 84.25 % compared to the AE and CNN models discussed earlier. The exhaustive review of relevant literature is elucidated in a tabulated format in Table.

The novel contribution of this research lies in its thorough exploration of the comparative performance of ensemble learning techniques using bagging-boosting and neural network architectures in network anomaly detection. While previous studies have examined individual models, this research bridges the gap by undertaking a holistic analysis that encompasses both traditional and cutting-edge methodologies. However, existing anomaly detection methods [12] often struggle to achieve high accuracy and robustness due to the evolving nature of network threats. This limitation can lead to missed attacks or false positives, hindering network security. Our proposed work addresses this challenge by proposing a novel ensemble model that leverages the strengths of established algorithms like XGBoost, Random Forest, Gradient Boosting, and Support Vector Machines. This hybrid approach aims to achieve superior accuracy and robustness in network anomaly detection. The emphasis on diverse model evaluation, trade-offs, and real-world applicability distinguishes this research from existing literature.

Experiments

This section elucidates the proposed ensemble framework designed for anomaly detection within the dataset. Commencing with a detailed exposition on the utilized dataset and its preprocessing procedures, the subsequent discussion delves into the distinctive contributions of each phase within the proposed framework, as illustrated in Fig. 1.

Dataset

The research harnesses the NSL KDD dataset, a pivotal benchmark in intrusion detection sourced from[1]. This dataset is refined and updated version of KDD Cup ’99

[1] Available at: https://www.unb.ca/cic/datasets/nsl.html (accessed: 19.03.2024).

Table. Review of Prominent work done in Network Anomaly Detection

References	Research Focus	Techniques Used	Dataset	Accuracy, %
[6]	Network traffic-based Anomaly detection	BLSTM and an attention method	NSL KDD	84.25
[9]	Anomaly Detection Framework using AE	AE	NSL KDD	90.61
[11]	Internet of the Things (IoT) based IDS	CNN	NSL KDD	99.40
[10]	Deep Learning based Network Intrusion Detection Model	Deep AE	KDD Cup ’99, NSL KDD	85 on the NSL KDD & 97.85 on the KDD Cup ’99
[13]	Neural network-based Intrusion Detection System (IDS)	Long Short-Term Memory (LSTM)-Recurrent Neural Networks (RNN)	KDD Cup ’99	93.82
[14]	LSTM RNN based IDS	LSTM-RNN	KDD Cup ’99	96.93
[15]	Network traffic-based anomaly detection	Ensemble of Principal Component Analysis (PCA) fuzzy based KNN	NSL KDD	98.24
[16]	Network Traffic analysis for anomaly detection	Clustering based framework used	UNSW-NB 15, CICIDS2017	97.90
[17]	Intrusion Detection for Internet of Medical Things	Random Forest with Grid Search	Manual dataset	94.23
[12]	Anomaly Detection in SDN	Ensemble Approach based on RandomForestClassifier, ExtraTreeClassifier, AdaBoostClassifier, GradientBoostingClassifier and XGBoost	NSL KDD	Around 99
[18]	Anomaly detection framework to detect network traffic	Linear SVM (Support Vector Machine), multilayer perceptron (MLP), Naive Bayes (NB)	Three different manual datasets	96.90
[19]	Model developed to detect anomaly in Network intrusion detection	Generative adversarial networks	CIC IDS 2017	82

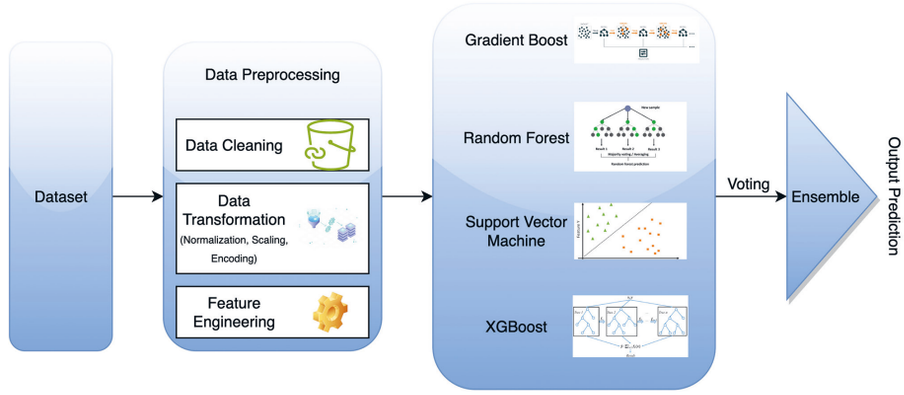

Fig. 1. Proposed ensemble model architecture

dataset [20], originally crafted by the Defense Advanced Research Projects Agency (DARPA), the NSL KDD dataset undergoes meticulous curation. Redundant and noisy data are expunged, and class balance is meticulously maintained, ensuring parity between normal and anomalous instances. The dataset contains two files namely training and test data. The dataset furnishes a rich substrate for analysis with 41 features encompassing network traffic intricacies, such as IP addresses, port numbers, and protocol types. It delineates between benign and malicious traffic as categorized into normal and attack facilitating nuanced anomaly detection.

Data preprocessing is executed meticulously to ensure the integrity and relevance of highly correlated features. The dataset under consideration amalgamates both training and test data, constituting a comprehensive repository of total 148,517 records. This amalgamation ensures a holistic representation of the data essential for training and evaluating the proposed ensemble-based anomaly detection model. Further, dataset is categorized into training and test data in the ratio of 60 % and 20 %, using train test split. An initial analytical overview provides a thorough examination of the dataset facilitating the categorization of attack types into distinct categories. The classification schema delineates attacks into five categories: Denial of Service (DoS), Probe, Remote-to-Local (R2L), User-to-Root (U2R), and Normal Traffic Data. This categorization framework underlies the subsequent anomaly detection methodologies to enable the model in perceiving and classifying diverse attack patterns. The standard preprocessing pipeline is further invoked here to fortify the dataset against inherent challenges and enhance its suitability for modeling. From correlation, the data features are sorted where by only features that have a higher correlation with each other were extracted. This is due to multifaceted approach commencing with handling of missing values in order data completeness. Numerical features are normalized and scaled, an indispensable step to make their scales in harmony and to contribute to model stability. One hot encoding is preferred instead of label encoding so as to prevent arbitrary order in categorical variables. One-hot encoding is the transformation process applied in subjecting categorical variables so that one can be able to represent the former in a form suitable for training machine learning models. Through this approach, information of categorical nature remains intact but at the same time becomes suitable with numerical-based algorithms. Another refinement involves the judicious elimination of irrelevant columns streamlining of the dataset to reduce such risk of complexities emanating from dimensionality. Further distillation of the dataset is performed through a feature selection endeavor and gave way to retaining the top 30 most highly correlated features. These features[1] capture the dataset essence that pertains to the salient information important in the anomaly detection task.

Ensemble Model Architecture

Using a rich tapestry of machine learning algorithms, ranging from XGBoost, Random Forest, Gradient Boosting, as well as SVM, the ensemble model is seamlessly woven into an integrated ensemble structure. The synergistic interplay of the models taps their respective architectural strengths, each contributing a unique vantage point in the intricate task of anomaly detection. These machine learning algorithms selected judiciously considering pros and cons of each model and tested them with Grid search algorithm. Gradient boosting methodology that XGBoost follows boosts its caliber to capture intricate patterns. Random Forest works against overfitting due to the use of decision trees. Gradient Boosting tunes for refining predictions sequentially, and SVM is known for its discriminative decision boundaries. This combination guarantees an almost exhaustive coverage of possible patterns in the data for variance in complexities.

The proposed ensemble approach for enhanced anomaly detection in network security leverages a comprehensive integration of bagging and boosting techniques, as illustrated in Fig. 1. The dataset undergoes a sequential passage through each model within the ensemble allowing for the exploitation of their individual strengths and capabilities. This strategic sequencing enables the ensemble to derive meaningful insights from the data while maximizing the detection of anomalous patterns. Central to our approach is the pre-training of the ensemble base classifier using the dataset. This initial training phase ensures that the ensemble is equipped with a foundational understanding of both normal and anomalous network behaviors. Once trained, the ensemble operates as a

[1] Available at: https://drive.google.com/file/d/13tahTGhvxb-1wzlefe3Nt1pMZnLkplbo/view (accessed: 19.03.2024).

cohesive unit, collectively analyzing incoming data to identify potential anomalies. Each base classifier within the ensemble evaluates the data independently, generating predictions that encompass both the predicted label and the posterior probability of samples. This holistic assessment enables the ensemble to capture a comprehensive view of the data landscape, enhancing its anomaly detection capabilities. The innovative aspect of our approach is the incorporation of a sophisticated weighted voting scheme which enhances the ensemble decision-making process. In order to determine the weight assigned to each model in the ensemble, we utilize a weighted voting mechanism based on accuracy performance metrics. The process begins with the evaluation of individual model accuracy on the training data. Each model is trained on the pre-processed training data, and its accuracy is assessed based on its ability to correctly classify data points as normal or anomalous. These accuracy scores serve as the basis for weight assignment, reflecting the relative performance of each model.

The weight assignment process takes into account both the priority of each model and its accuracy performance. Priority is assigned based on the significance of each model contribution to the ensemble, with higher priority models receiving greater weights. For example, XGBoost, with its superior ability to capture complex patterns, may be assigned the highest weight, followed by Gradient Boosting, Random Forest, and Support Vector Classifier (SVC). The weights are assigned proportionally to the accuracy scores ensuring that models with higher accuracy contribute more significantly to the final decision-making process. The justification for this weight assignment scheme lies in the balance between accuracy and the unique characteristics of each model. Models like XGBoost and Gradient Boosting, with their ability to handle complex patterns, are likely to achieve higher accuracy scores. Random Forest, while potentially having slightly lower accuracy, plays a crucial role in preventing overfitting and improving generalization. SVC, with its interpretable decision boundaries, have slightly lower accuracy compared to other models but still contributes valuable insights to the ensemble. The ensemble model operates in a sequential manner, with the results of each base classifier cascading into the subsequent voting mechanism. This iterative process not only ensures comprehensive data analysis but also effectively neutralizes the idiosyncrasies associated with individual models. Moreover, the ensemble adopts a soft voting mechanism, departing from rigid binary voting strategies. This nuanced decision-making process takes into account the confidence levels of each individual model, thereby enhancing adaptability to uncertain or ambiguous scenarios.

Once weights are assigned to each model, the weighted voting process is employed to combine the predictions of individual models and make the final decision on anomaly detection. For each data point in the test set, the predicted class label (1 for normal, 0 for anomaly) from each model is multiplied by its corresponding weight. The weighted votes across all models are then summed, and the sum is compared to a predefined threshold (e.g., 0.5). If the sum exceeds or equals the threshold, the final ensemble prediction is labeled as “normal”; otherwise, it’s classified as “anomaly”.

A Comprehensive Ensemble Approach Integrating Bagging and Boosting Techniques

Our proposed ensemble approach implementation requires a rigorous process that allows each individual model in the ensemble to deliver optimal performance. We used the grid search algorithm with a more sophisticated method of hyperparameters tuning. This method scans through each defined space of hyperparameters extensively, thus covering all possible configurations for each of the models making up the ensemble in a systematic manner. The grid search algorithm tooks nearly 12 hours to get the final parameters that were used in this study. The settings of hyperparameters for individual models are carefully selected to balance between complexity and performance. The Gradient Boosting Classifier, Random Forest Classifier, SVC (with a linear kernel and $C = 0.1$), and XGBoost Classifier serve as the foundational models within our ensemble. Their parameter configurations are fine-tuned to optimize their individual contributions. We set the number of trees (*n_estimators*) to 200 for both Gradient Boosting and Random Forest algorithms. This parameter controls the complexity of the ensemble model by determining the number of decision trees used in the final prediction. Additionally, the maximum depth (*max_depth*) of each tree is set to 20 for both algorithms. This parameter limits the maximum number of splits allowed in each tree, preventing them from becoming overly complex and potentially overfitting the training data. For the SVC, the regularization parameter (*C*) is set to 0.1. This parameter controls the trade-off between fitting the training data and keeping the model generalizable. Finally, the XGBoost Classifier utilizes a combination of *n_estimators* (200), *max_depth* (6), and *learning_rate* (0.3). The *learning_rate* parameter controls the step size taken when the model updates its internal parameters during training.

In preparation for training, the ensemble model was trained on pre-processed training dataset whereby all models learn diverse patterns that describe both normal and anomalous network behaviors. The intention is to make complex knowledge of the data set available to the ensemble so as to improve its anomaly detection ability. After that, we have thoroughly evaluated how our model performs using a test set which serves as an important evaluation criterion for its generalization capability. This evaluation phase involves assessing the model ability to extrapolate its learned patterns to unseen data.

Experimental Setup

The experimental setup for this research leveraged Google Collaboratory, a cloud-based Integrated Development Environment (IDE), for model execution. The utilization of a robust GPU within Google Collaboratory substantially reduced the training time for the proposed ensemble model. The values essential for computations were derived from the Google Compute Engine backend specifically utilizing a GPU. The system RAM usage reached 2.7 GB, representing a fraction of the total 12.7 GB capacity. The GPU, identified as an NVIDIA Tesla T4, played a pivotal role in enhancing computational efficiency. Furthermore, the disk space allocation amounted

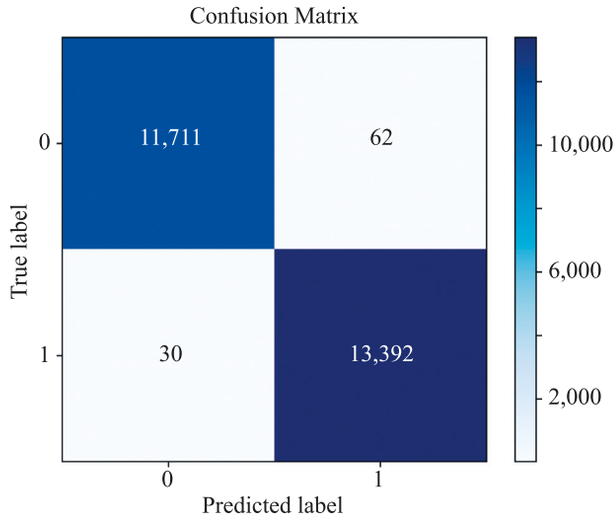

Fig. 2. Confusion matrix of proposed model

to 21.3 GB out of a total capacity of 78.2 GB, ensuring ample storage for the seamless execution of the research.

Results

Performance Evaluation Metrics

The evaluation of our ensemble model was carried out with the full suite of performance metrics. The metrics that we considered for the evaluation of our research are accuracy, precision, recall, F1-score and Area Under the Receiver Operating Characteristic curve (AUC-ROC). These we used together to detect anomaly accurately while reducing false positive and false negatives.

Accuracy is a fundamental metric to find the overall correctness of predictions. Hence precision becomes crucial when considering false positives cost measuring as an anomaly equivalent model accuracy. Moreover, recall measures how much the model is able to discover all the actual anomalies, pointing out the true positives ratio among all positive actual cases. F1-score chose to combine precision and recall in order to provide a measure balanced, extremely valuable especially for imbalanced datasets. The AUC-ROC is crucial in the analysis of performance associated with binary classification tasks. Actually, it is one of the performance measures and hence forms a foundation upon whose foundation other performance measures are established against. The ROC curve plots the relation between true positive rate and false positive rate at various decision thresholds. AUC-ROC quantifies the discriminatory ability of the model where the better model discriminating normal and anomalous instances tends to have higher values. Also, log loss depicts the overall loss during the training of the model and guarantees the correctness of the model. Together, these metrics collectively produce a nuanced evaluation allowing comprehensively understand strengths and limitations of an anomaly detection model [21].

The accuracy of proposed model showcase its proficiency, boasting a remarkable achievement of 99.68 %. This numerical testament underscores the model adeptness in distinguishing between normal and anomalous instances within the dataset. Fig. 3, *a* provided a holistic view of the model overall predictive correctness, portraying a trajectory of consistently high accuracy across different scenarios. The confusion matrix plot as depicted in Fig. 2 dissected the model classification outcomes, delineating true positives, false positives, true negatives, and false negatives with granular precision. The model extends its performance in terms of precision, F1-score and recall as 99.54, 99.66 and 99.78 as shown in Fig. 3, *b*, Fig. 4, *a* and Fig. 4, *b* respectively. On the other hand, the model ability to balance precision and recall, encapsulated by the F1-score plot, attested to its equilibrium in minimizing false positives and false negatives is shown in Fig. 4, *a*. The Log Loss plot furnished a nuanced depiction of the model calibration, capturing the intricacies of probability estimation. The Log Loss of model showing the training loss over base classifiers are depicted in Fig. 5, *a*.

Precision and recall, vital cogs in anomaly detection, were graphically articulated through dedicated precision and recall plots. These visual aids elucidated the trade-off between accurately identified anomalies and potential false positives. The quintessential AUC-ROC plots encapsulated the model discriminatory power across varying thresholds, offering a comprehensive view of its true positive rate against false positive rate is depicted in Fig. 5, *b*.

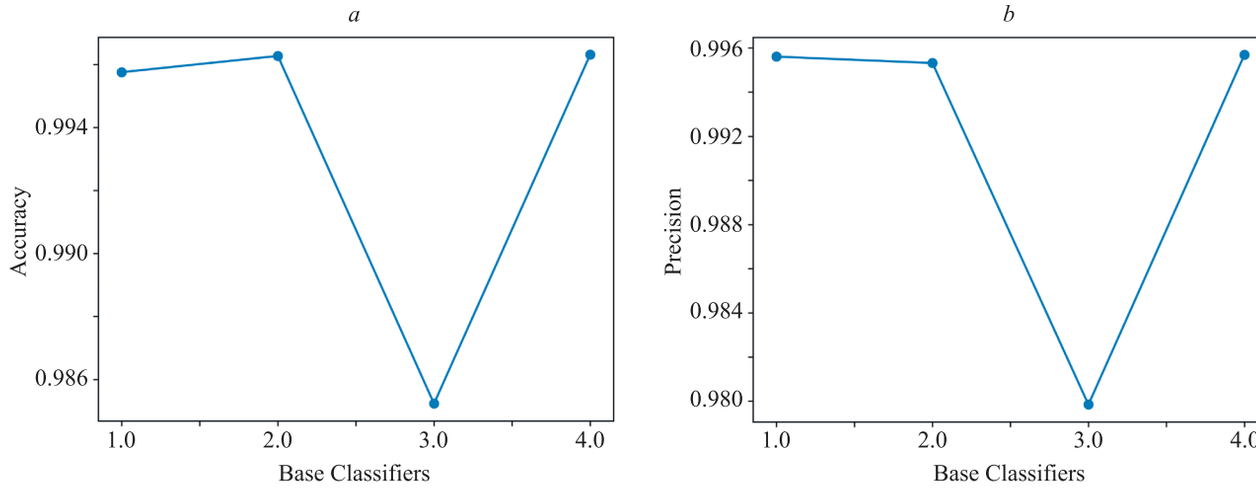

Fig. 3. Accuracy (*a*) and Precision (*b*) vs. base classifiers in the proposed ensemble approach

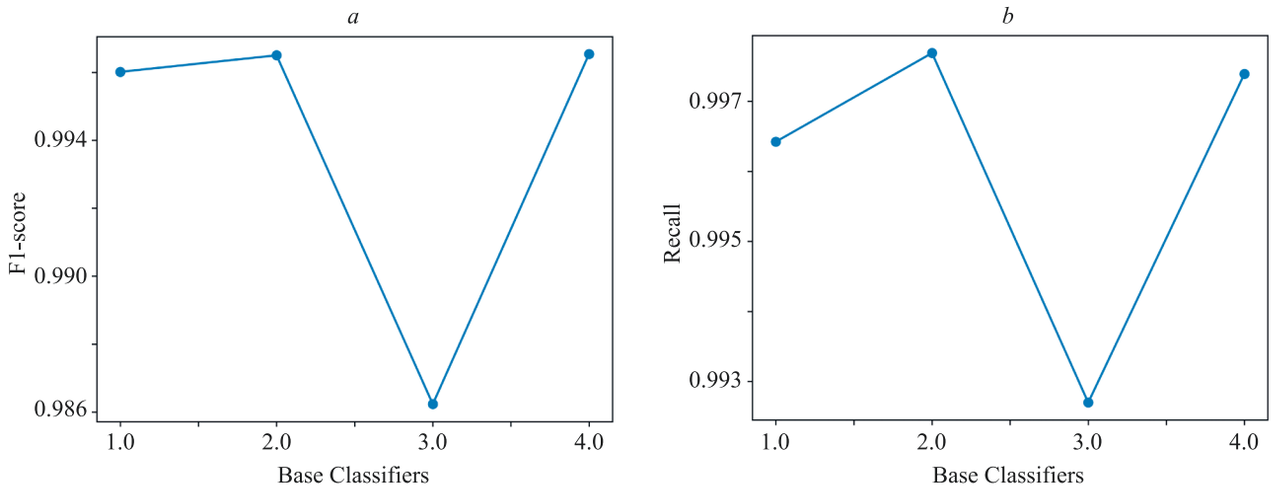

Fig. 4. F1-score (*a*) and Recall (*b*) vs. base classifiers in the proposed ensemble approach

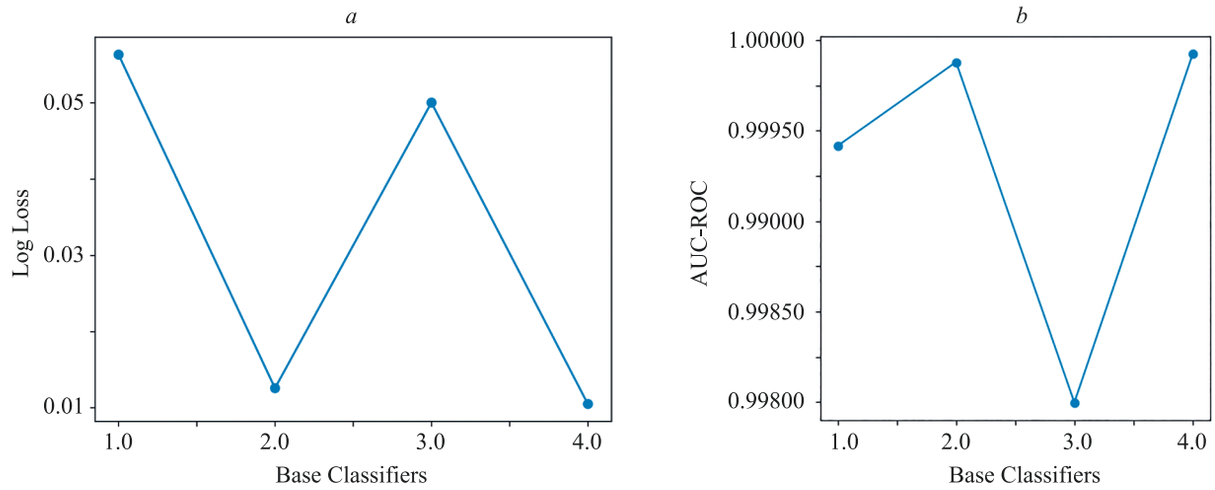

Fig. 5. Log Loss (*a*) and AUC-ROC (*b*) vs. base classifiers in the proposed ensemble approach

Comparative Study

We take into account the previous work from [9–11, 22–25] to compare our research. Our model achieves an accuracy of 99.68 %, far better than existing [9] and [11] benchmarks. In addition, our proposed ensemble models excel in precision, recall and F1-score, demonstrated by its sturdiness across a variety of indicators. Notably, in comparison with the methods noted in [22–25], the proposed framework showed a degree of superiority. This shows that it is adept at dealing with some of the mettle test entailed by anomaly detection. The nuanced comparison underscores the significance of our proposed ensemble model in advancing the state-of-the-art in anomaly detection.

Conclusion

In conclusion, the study introduces a solid ensemble framework for the anomaly detection attempts exploiting the fusion of different base classifiers within a complex voting mechanism. The careful choice of dataset and individual models in ensemble approach commences the building of a holistic analysis. The Gradient Boosting, Random Forest, Support Vector Machine, and XGBoost brought as base classifiers that enhance the discernment of ensemble capture the subtler patterns on aggregation. The presented ensemble is one capable model described through a rigorous evaluation with key metrics and produced effective results. The theatrics show that the proposed framework is able to capture anomalies within the dataset very effectively with an accuracy of 99.68 %. The binary classification indicative of anomaly presence for the ensemble indicates the offered practical utility of the ensemble in real world scenarios.

Even though this contributes quite a lot to network anomaly detection, our research has its own limitations. Here are some of them. First, the study is delimited to a specific dataset that may impact making generalization between this result and those that exist in other network environment. Lack of real-time experimentation and elimination of hybrid models preclude dynamic adaptations that are bound to evolving threats. Stronger understanding of adversarial attacks and concurrent deployment is an essential booster to the resilience in our model. This is further the temporal characteristic of network behavior that could best be served by temporal analysis and online learning methodologies.

References

1. Reichenbach M. New challenges in electronic payments. *Intelligent Enterprises of the 21st Century*. Ed. by J.N.D. Gupta, S. Sharma. IGI Global, 2004, pp. 153–162. https://doi.org/10.4018/9781591401605.ch010
2. Kebande V.R., Karie N.M., Ikuesan R.A. Real-time monitoring as a supplementary security component of vigilantism in modern network environments. *International Journal of Information Technology*, 2021, vol. 13, no. 1, pp. 5–17. https://doi.org/10.1007/s41870-020-00585-8
3. Hareesh R., Senthil Kumar R.K., Kalluri R., Bindhumadhava B.S. Critical infrastructure asset discovery and monitoring for cyber security. *Lecture Notes in Electrical Engineering*, 2022, vol. 847, pp. 289–300. https://doi.org/10.1007/978-981-16-9008-2_27
4. Savage D., Zhang X., Yu X., Chou P., Wang Q. Anomaly detection in online social networks. *Social Networks*, 2014, vol. 39, pp. 62–70. https://doi.org/10.1016/j.socnet.2014.05.002
5. Benaddi H., Ibrahimi K., Benslimane A. Improving the intrusion detection system for NSL-KDD dataset based on PCA-fuzzy clustering-KNN. *Proc. of the 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, 2018, pp. 1–6. https://doi.org/10.1109/wincom.2018.8629718
6. Su T., Sun H., Zhu J., Wang S., Li Y. BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset. *IEEE Access*, 2020, vol. 8, pp. 29575–29585. https://doi.org/10.1109/access.2020.2972627
7. Wang C., Zhou H., Hao Z., Hu S., Li J., Zhang X., Jiang B., Chen X. Network traffic analysis over clustering-based collective anomaly detection. *Computer Networks*, 2022, vol. 205, pp. 108760. https://doi.org/10.1016/j.comnet.2022.108760
8. Keim Y., Mohapatra A.K. Cyber threat intelligence framework using advanced malware forensics. *International Journal of Information Technology*, 2022, vol. 14, no. 1, pp. 521–530. https://doi.org/10.1007/s41870-019-00280-3
9. Xu B., Jang-Jaccard J., Singh A., Wei Y., Sabrina F. Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset. *IEEE Access*, 2021, vol. 9, pp. 140136–140146. https://doi.org/10.1109/access.2021.3116612
10. Shone N., Ngoc T.N., Phai V.D., Shi Q. A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, vol. 2, no. 1, pp. 41–50. https://doi.org/10.1109/tetci.2017.2772792
11. Sharma B., Sharma L., Lal C. Anomaly based network intrusion detection for IoT attacks using convolution neural network. *Proc. of the IEEE 7th International Conference for Convergence in Technology (I2CT)*, 2022, pp. 1–6. https://doi.org/10.1109/i2ct54291.2022.9824229
12. Krzemień W., Jędrasiak K., Nawrat A. Anomaly detection in software defined networks using ensemble learning. *Lecture Notes in Networks and Systems*, 2022, vol. 439, pp. 629–643. https://doi.org/10.1007/978-3-030-98015-3_44
13. Staudemeyer R.C. Applying long short-term memory recurrent neural networks to intrusion detection. *South African Computer Journal*, 2015, vol. 56, pp. 136–154. https://doi.org/10.18489/sacj.v56i1.248
14. Kim J., Kim J., Thu H.L.T., Kim H. Long short term memory recurrent neural network classifier for intrusion detection. *Proc. of the International Conference on Platform Technology and Service (PlatCon)*, 2016, pp. 1–5. https://doi.org/10.1109/platcon.2016.7456805
15. Liu Z., Thapa N., Shaver A., Roy K., Yuan X., Khorsandroo S. Anomaly detection on IoT network intrusion using machine learning. *Proc. of the International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (ICABCD)*, 2020, pp. 1–5. https://doi.org/10.1109/icabcd49160.2020.9183842
16. Khan W., Haroon M. An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks. *International Journal of Cognitive Computing in Engineering*, 2022, vol. 3, pp. 153–160. https://doi.org/10.1016/j.ijcce.2022.08.002
17. Gupta K., Sharma D.K., Gupta K.D., Kumar A. A tree classifier based network intrusion detection model for Internet of Medical Things. *Computers and Electrical Engineering*, 2022, vol. 102, pp. 108158. https://doi.org/10.1016/j.compeleceng.2022.108158
18. Ma Q., Sun C., Cui B. A novel model for anomaly detection in network traffic based on support vector machine and clustering. *Security and Communication Networks*, 2021, pp. 170788. https://doi.org/10.1155/2021/2170788

Литература

1. Reichenbach M. New challenges in electronic payments // Intelligent Enterprises of the 21st Century / ed. by J.N.D. Gupta, S. Sharma. IGI Global, 2004. P. 153–162. https://doi.org/10.4018/9781591401605.ch010
2. Kebande V.R., Karie N.M., Ikuesan R.A. Real-time monitoring as a supplementary security component of vigilantism in modern network environments // International Journal of Information Technology. 2021. V. 13. N 1. P. 5–17. https://doi.org/10.1007/s41870-020-00585-8
3. Hareesh R., Senthil Kumar R.K., Kalluri R., Bindhumadhava B.S. Critical infrastructure asset discovery and monitoring for cyber security // Lecture Notes in Electrical Engineering. 2022. V. 847. P. 289–300. https://doi.org/10.1007/978-981-16-9008-2_27
4. Savage D., Zhang X., Yu X., Chou P., Wang Q. Anomaly detection in online social networks // Social Networks. 2014. V. 39. P. 62–70. https://doi.org/10.1016/j.socnet.2014.05.002
5. Benaddi H., Ibrahimi K., Benslimane A. Improving the intrusion detection system for NSL-KDD dataset based on PCA-fuzzy clustering-KNN // Proc. of the 6th International Conference on Wireless Networks and Mobile Communications (WINCOM). 2018. P. 1–6. https://doi.org/10.1109/wincom.2018.8629718
6. Su T., Sun H., Zhu J., Wang S., Li Y. BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset // IEEE Access. 2020. V. 8. P. 29575–29585. https://doi.org/10.1109/access.2020.2972627
7. Wang C., Zhou H., Hao Z., Hu S., Li J., Zhang X., Jiang B., Chen X. Network traffic analysis over clustering-based collective anomaly detection // Computer Networks. 2022. V. 205. P. 108760. https://doi.org/10.1016/j.comnet.2022.108760
8. Keim Y., Mohapatra A.K. Cyber threat intelligence framework using advanced malware forensics // International Journal of Information Technology. 2022. V. 14. N 1. P. 521–530. https://doi.org/10.1007/s41870-019-00280-3
9. Xu B., Jang-Jaccard J., Singh A., Wei Y., Sabrina F. Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset // IEEE Access. 2021. V. 9. P. 140136–140146. https://doi.org/10.1109/access.2021.3116612
10. Shone N., Ngoc T.N., Phai V.D., Shi Q. A deep learning approach to network intrusion detection // IEEE Transactions on Emerging Topics in Computational Intelligence. 2018. V. 2. N 1. P. 41–50. https://doi.org/10.1109/tetci.2017.2772792
11. Sharma B., Sharma L., Lal C. Anomaly based network intrusion detection for IoT attacks using convolution neural network // Proc. of the IEEE 7th International Conference for Convergence in Technology (I2CT). 2022. P. 1–6. https://doi.org/10.1109/i2ct54291.2022.9824229
12. Krzemień W., Jędrasiak K., Nawrat A. Anomaly detection in software defined networks using ensemble learning // Lecture Notes in Networks and Systems. 2022. V. 439. P. 629–643. https://doi.org/10.1007/978-3-030-98015-3_44
13. Staudemeyer R.C. Applying long short-term memory recurrent neural networks to intrusion detection // South African Computer Journal. 2015. V. 56. P. 136–154. https://doi.org/10.18489/sacj.v56i1.248
14. Kim J., Kim J., Thu H.L.T., Kim H. Long short term memory recurrent neural network classifier for intrusion detection // Proc. of the International Conference on Platform Technology and Service (PlatCon). 2016. P. 1–5. https://doi.org/10.1109/platcon.2016.7456805
15. Liu Z., Thapa N., Shaver A., Roy K., Yuan X., Khorsandroo S. Anomaly detection on IoT network intrusion using machine learning // Proc. of the International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (ICABCD). 2020. P. 1–5. https://doi.org/10.1109/icabcd49160.2020.9183842
16. Khan W., Haroon M. An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks // International Journal of Cognitive Computing in Engineering. 2022. V. 3. P. 153–160. https://doi.org/10.1016/j.ijcce.2022.08.002
17. Gupta K., Sharma D.K., Gupta K.D., Kumar A. A tree classifier based network intrusion detection model for Internet of Medical Things // Computers and Electrical Engineering. 2022. V. 102. P. 108158. https://doi.org/10.1016/j.compeleceng.2022.108158
18. Ma Q., Sun C., Cui B. A novel model for anomaly detection in network traffic based on support vector machine and clustering // Security and Communication Networks. 2021. P. 170788. https://doi.org/10.1155/2021/2170788
19. Iliyasu A.S., Deng H. N-GAN: a novel anomaly-based network intrusion detection with generative adversarial networks //

19. Iliyasu A.S., Deng H. N-GAN: a novel anomaly-based network intrusion detection with generative adversarial networks. *International Journal of Information Technology*, 2022, vol. 14, no. 7, pp. 3365–3375. https://doi.org/10.1007/s41870-022-00910-3
20. Tavallaee M., Bagheri E., Lu W., Ghorbani A.A. A detailed analysis of the KDD CUP 99 data set. *Proc. of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6. https://doi.org/10.1109/cisda.2009.5356528
21. Panesar A. Evaluating machine learning models. *Machine Learning and AI for Healthcare*. Apress, Berkeley, CA, 2021, pp. 189–205. https://doi.org/10.1007/978-1-4842-6537-6_7
22. Assy A.T., Mostafa Y., Abd El-khaleq A., Mashaly M. Anomaly-based intrusion detection system using one-dimensional convolutional neural network. *Procedia Computer Science*, 2023, vol. 220, pp. 78–85. https://doi.org/10.1016/j.procs.2023.03.013
23. Acharya T., Annamalai A., Chouikha M.F. Efficacy of bidirectional LSTM model for network-based anomaly detection. *Proc. of the IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2023, pp. 336–341. https://doi.org/10.1109/iscaie57739.2023.10165336
24. Kavitha S., Uma Maheswari N., Venkatesh R. Network anomaly detection for NSL-KDD dataset using deep learning. *Information Technology in Industry*, 2021, vol. 9, no. 2, pp. 821–827. https://doi.org/10.17762/itii.v9i2.419
25. Gadal S., Mokhtar R., Abdelhaq M., Alsaqour R., Ali E.S., Saeed R. Machine learning-based anomaly detection using K-mean array and sequential minimal optimization. *Electronics*, 2022, vol. 11, no. 14, pp. 2158. https://doi.org/10.3390/electronics11142158

International Journal of Information Technology. 2022. V. 14. N 7. P. 3365–3375. https://doi.org/10.1007/s41870-022-00910-3
20. Tavallaee M., Bagheri E., Lu W., Ghorbani A.A. A detailed analysis of the KDD CUP 99 data set // Proc. of the IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2009. P. 1–6. https://doi.org/10.1109/cisda.2009.5356528
21. Panesar A. Evaluating machine learning models // Machine Learning and AI for Healthcare. Apress, Berkeley, CA, 2021. P. 189–205. https://doi.org/10.1007/978-1-4842-6537-6_7
22. Assy A.T., Mostafa Y., Abd El-khaleq A., Mashaly M. Anomaly-based intrusion detection system using one-dimensional convolutional neural network // Procedia Computer Science. 2023. V. 220. P. 78–85. https://doi.org/10.1016/j.procs.2023.03.013
23. Acharya T., Annamalai A., Chouikha M.F. Efficacy of bidirectional LSTM model for network-based anomaly detection // Proc. of the IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE). 2023. P. 336–341. https://doi.org/10.1109/iscaie57739.2023.10165336
24. Kavitha S., Uma Maheswari N., Venkatesh R. Network anomaly detection for NSL-KDD dataset using deep learning // Information Technology in Industry. 2021. V. 9. N 2. P. 821–827. https://doi.org/10.17762/itii.v9i2.419
25. Gadal S., Mokhtar R., Abdelhaq M., Alsaqour R., Ali E.S., Saeed R. Machine learning-based anomaly detection using K-mean array and sequential minimal optimization // Electronics. 2022. V. 11. N 14. P. 2158. https://doi.org/10.3390/electronics11142158

Authors

Rashmikiran Pandey — PhD Student, Moscow Institute of Physics and Technology (National Research University), Moscow region, Dolgoprudny, 141701, Russian Federation, SC 58122041300, https://orcid.org/0000-0003-0042-6565, rashmikiran@phystech.edu

Mrinal Pandey — PhD Student, Moscow Institute of Physics and Technology (National Research University), Moscow region, Dolgoprudny, 141701, Russian Federation, SC 58425838700, https://orcid.org/0009-0009-5151-6908, mrinalpandei@phystech.edu

Alexey N. Nazarov — D.Sc., Professor, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333, Russian Federation, SC 7201780424, https://orcid.org/0000-0002-0497-0296, a.nazarov06@bk.ru

Авторы

Пандей Рашмикиран — аспирант, Московский физико-технический институт (национальный исследовательский университет), Московская область, Долгопрудный, 141701, Российская Федерация, SC 58122041300, https://orcid.org/0000-0003-0042-6565, rashmikiran@phystech.edu

Пандей Мринал — аспирант, Московский физико-технический институт (национальный исследовательский университет), Московская область, Долгопрудный, 141701, Российская Федерация, SC 58425838700, https://orcid.org/0009-0009-5151-6908, mrinalpandei@phystech.edu

Назаров Алексей Николаевич — доктор технических наук, профессор, Федеральный исследовательский центр «Информатика и управление» Российской академии наук», Москва, 119333, Российская Федерация, SC 7201780424, https://orcid.org/0000-0002-0497-0296, a.nazarov06@bk.ru

Received 08.01.2024
Approved after reviewing 26.07.2024
Accepted 16.09.2024

Статья поступила в редакцию 08.01.2024
Одобрена после рецензирования 26.07.2024
Принята к печати 16.09.2024